

ANÁLISE DE REGRESSÃO LINEAR SIMPLES COM ERRO NA VARIÁVEL INDEPENDENTE

Hao Min Huai e Sarita Papescu

Bolsistas do CNPq

Orientadora: Lisbeth Kaiserlian Cordani

Muitas vezes ao utilizarmos a análise de Regressão Linear Simples, nos deparamos com a variável independente aleatória sujeita a erros de mensuração. Este trabalho tem por objetivo indicar estimadores dos parâmetros do modelo de regressão quando estamos diante de uma situação como esta.

I. Modelo

Considere a relação:

$$y = \beta_0 + \beta_1 x \quad (1)$$

Dados n pares de valores (x_i, y_i) , suponhamos que não podemos observar os “verdadeiros” valores de x e y , mas sim os valores de duas variáveis aleatórias X e Y , tais que:

$$\begin{cases} Y_i = y_i + \varepsilon_i \\ X_i = x_i + u_i, \quad i = 1, 2, \dots, n. \end{cases} \quad (2)$$

onde ε_i e u_i são erros aleatórios.

Vamos supor que:

$$\varepsilon_i \sim N(0, \sigma_\varepsilon^2), \quad \forall i$$

$$u_i \sim N(0, \sigma_u^2), \quad \forall i \quad \text{independentes identicamente distribuídos}$$

$$x_i \sim N(u_x, \sigma_x^2), \quad \forall i$$

Substituindo (2) em (1), temos:

$$Y_i = \beta_0 + \beta_1 X_i + v_i \quad (3)$$

onde $v_i = \varepsilon_i - \beta_1 u_i$.

O modelo em (3) tem a aparência de um modelo clássico de regressão simples com erro v_i , mas uma suposição preliminar daquela análise que não é obedecida aqui é que a covariância entre X e o erro associado a Y é igual a zero. De fato,

$$\text{cov}(X_i, v_i) = E(X_i v_i) - E(X_i) \cdot E(v_i) = -\beta_1 \sigma_u^2 \quad (4)$$

que será zero apenas nos casos em que $\sigma_u^2 = 0$ (regressão clássica) ou $\beta_1 = 0$.

Se fôssemos estimar β_1 pelo método usual de mínimos quadrados ordinários (MQO) teríamos

$$\hat{\beta}_{\text{MQO}} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \beta_1 + \frac{\sum (X_i - \bar{X})v_i}{\sum (X_i - \bar{X})^2} \quad (5)$$

que é um estimador viciado de β_1 , pois $E(\hat{\beta}_{\text{MQO}}) \neq \beta_1$.

Além disso,

$$\hat{\beta}_{\text{MQO}} \xrightarrow{p} \beta_1 + \frac{-\beta_1 \sigma_u^2}{\sigma_x^2 + \sigma_u^2} = \frac{\beta_1}{1 + (\sigma_u^2/\sigma_x^2)} \leq \beta_1 \quad (6)$$

Logo, $\hat{\beta}_{\text{MQO}}$ é inconsistente e subestima o verdadeiro valor do parâmetro. O vício do estimador é dado por

$$K = \frac{1}{1 + (\sigma_u^2/\sigma_x^2)} = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} \leq 1 \quad (7)$$

Esse termo é conhecido na literatura como Coeficiente de Atenuação.

Então, a solução tradicional para estimação de β_1 não é adequada.

II. Problema de Identificação

O principal problema na estimação dos parâmetros neste contexto é chamado "Problema de Identificação", que mencionaremos a seguir.

(X, Y) tem distribuição Normal Bivariada, ou seja:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N_2 \left(\begin{bmatrix} \mu_x \\ \beta_0 + \beta_1 \mu_x \end{bmatrix}; \begin{bmatrix} \sigma_u^2 + \sigma_\varepsilon^2 & \beta_1 \sigma_\varepsilon^2 \\ \beta_1 \sigma_\varepsilon^2 & \beta_1^2 \sigma_\varepsilon^2 + \sigma_\varepsilon^2 \end{bmatrix} \right)$$

Dada uma amostra de n pares (X_i, Y_i) , o conjunto de estatísticas suficientes associado a (X, Y) é dado por:

$$\bar{X}, \bar{Y}, s_X^2, s_Y^2 \text{ e } s_{XY}.$$

Podemos usar o método dos Momentos que, nesse caso, é equivalente ao de Máxima Verossimilhança, para estimar os parâmetros conhecidos: $\beta_0, \beta_1, \mu_x, \sigma_x^2, \sigma_e^2$ e σ_u^2 . Assim, obtemos o seguinte sistema de equações:

$$\begin{aligned} (1) \quad \hat{\mu}_x &= \bar{X} \\ (2) \quad \hat{\beta}_0 + \hat{\beta}_1 \hat{\mu}_x &= \bar{Y} \\ (3) \quad \hat{\sigma}_x^2 + \hat{\sigma}_u^2 &= s_X^2 \\ (4) \quad \hat{\beta}_1^2 \hat{\sigma}_x^2 + \hat{\sigma}_e^2 &= s_Y^2 \\ (5) \quad \hat{\beta}_1^2 \hat{\sigma}_x^2 &= s_{XY}. \end{aligned} \tag{8}$$

Diante de um sistema de 5 equações não conseguimos estimar 6 parâmetros. Esse problema é o chamado "Problema de Identificação". Uma solução seria obtermos alguma informação adicional a respeito dos parâmetros e nas seções seguintes analisaremos dois casos particulares.

III. Estimação de β_1 supondo conhecido o termo σ_x^2/σ_X^2

Como foi visto em (6) e (7)

$$\hat{\beta}_{MQO} \xrightarrow{p} \beta_1 \cdot K, \quad 0 < K \leq 1$$

Desta forma, um estimador de β_1 do modelo (3) pode ser dado por

$$\hat{\beta}_1 = \frac{\hat{\beta}_{MQO}}{K} = \frac{s_{xy}}{s_X^2 \cdot K} \tag{9}$$

Prova-se que $\hat{\beta}_1$, além de consistente, é não viciado e que

$$\text{Var}(\hat{\beta}_1) = \frac{\beta_1^2 \sigma_x^2 (1 - K) + \sigma_e^2}{K^2 (n - 3) (\sigma_x^2 + \sigma_u^2)} \quad (\text{ver [2]})$$

IV. Estimação de β_1 supondo conhecido o termo $\lambda = \frac{\sigma_z^2}{\sigma_u^2}$

Fazendo $\hat{\sigma}_\varepsilon^2 = \lambda \hat{\sigma}_u^2$ e substituindo em (8), temos

$$\hat{\beta}_1^2 s_{XY} + \hat{\beta}_1 (\lambda s_Y^2 - s_Y^2) - \lambda s_{XY} = 0 \quad (10)$$

Para $s_{XY} = 0$:

se $\frac{s_Y^2}{s_X^2} = \lambda$ então $\hat{\beta}_1$ é indeterminado;

se $\frac{s_Y^2}{s_X^2} \neq \lambda$ então $\hat{\beta}_1 = 0$.

Se $s_{XY} \neq 0$, então as soluções de (8) serão dadas por

$$\hat{\beta}_1 = \frac{(s_Y^2 - \lambda s_X^2) \pm \sqrt{(s_Y^2 - \lambda s_X^2)^2 + 4\lambda s_{XY}^2}}{2s_{XY}} = \frac{N}{2s_{XY}}$$

mas

$$\hat{\sigma}_\varepsilon^2 = \frac{s_{XY}}{\hat{\beta}_1} = \frac{2s_{XY}^2}{N}.$$

Logo, para $\hat{\sigma}_\varepsilon^2 \geq 0$, N deve assumir valores positivos, então

$$\hat{\beta}_1 = \frac{(s_Y^2 - \lambda s_X^2) + \sqrt{(s_Y^2 - \lambda s_X^2)^2 + 4\lambda s_{XY}^2}}{2s_{XY}} \quad (11)$$

isto é, tomamos somente a raiz positiva para calcular $\hat{\beta}_1$.

Verifica-se que os valores assumidos por $\hat{\sigma}_\varepsilon^2$ e $\hat{\sigma}_u^2$ são, de fato, não negativos para qualquer valor de $\hat{\beta}_1$ em (11).

O estimador de β_1 dado por (11) é o de Máxima Verossimilhança quando se conhece o termo $\lambda = \frac{\sigma_z^2}{\sigma_u^2}$, e sua variância é dada por

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma_\varepsilon^2 \sigma_u^2 + \beta_1^2 \sigma_u^2 \sigma_\varepsilon^2 + \sigma_\varepsilon^2 \sigma_u^2}{n \sigma_\varepsilon^4} \quad (\text{ver [2]})$$

V. Simulação

Foram feitas simulações por computador (utilizando PASCAL) para estudar o comportamento dos estimadores $\hat{\beta}_1$ (dados por (9) e por (11)) para vários valores de σ_ε^2 .

Para se medir a variabilidade de $\hat{\beta}_1$ em relação a β_1 , foi adotado o erro quadrático médio.

Resumidamente, os resultados indicaram que quanto maior a variância de $x(\sigma_x^2)$, menor é o erro quadrático médio de $\hat{\beta}_1$, isto é, melhor é a estimativa de β_1 . Além disso, quanto maior for o tamanho amostral utilizado para estimar β_1 , menor é seu erro quadrático médio. Em trabalhos posteriores serão comentadas as simulações com variação de σ_e^2 e σ_u^2 .

Bibliografia

- [1] Draper, N.R. and Smith, H. (1981). *Applied Regression Analysis*. John Wiley, 2nd. ed., New York.
- [2] Fuller, W.A. (1987). *Measurement Error Models*. John Wiley.
- [3] Johnston, J. (1963). *Econometrics Methods*. McGraw-Hill, New York.
- [4] Kendall, M.G. and Stuart, A. (1973). *The Advanced Theory of Statistics*, vol.2, 3rd. ed., London.